EQUALLY ACCURATE, UNEQUALLY LEGIBLE: MEASURING THE LEGIBILITY OF LLM ADJUDICATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used to resolve real disputes, from e-commerce claims to prediction-market arbitration. An adjudicator is chosen when a contract is signed, long before the disputes it will rule on exist. At that moment, the only evidence about a judge is its record, usually summarized by reference agreement: how often its rulings match those of the institution it would replace (e.g., a human jury). This tells a party how closely the judge tracks that institution, but not whether the party can anticipate its rulings. Following the legal-realist view of law as the prediction of what judges will do, we introduce adjudicative legibility: how well an external forecaster can predict an adjudicator's ruling on an unseen dispute, given the dispute and, optionally, its published rules and past rulings. We study eight LLM adjudicators on contested e-commerce and article-deletion disputes. On more than a third of them, the outcome depends on which adjudicator is chosen, and adjudicators of similar accuracy differ by up to $2.35\times$ in illegibility. We prove that a party can protect itself by demanding a safety margin that grows with a judge's illegibility; on these disputes, this margin is comparable to that with no forecast at all. A party cannot meaningfully consent to a judge it cannot anticipate; we therefore argue that as LLMs take on binding adjudicative roles, legibility should be measured as routinely as accuracy.

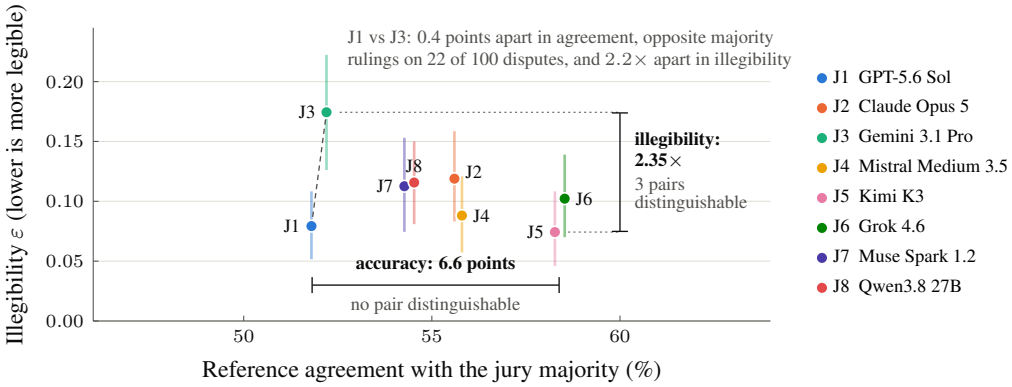

Figure 1: **Accuracy cannot choose a machine adjudicator.** Eight LLM adjudicators on the same 100 contested e-commerce disputes, placed by agreement with the jury majority (x) and by an observer's forecast error beyond each adjudicator's own randomness, given the dispute alone (y; 95% intervals). With intervals that hold jointly for all 28 pairs, no pair of adjudicators can be distinguished by agreement, while three can by legibility. For instance, J1 and J3 are 0.4 points apart in agreement but differ $2.2\times$ in illegibility.

1 INTRODUCTION

Large Language Models (LLMs) are increasingly being used to adjudicate between competing claims. Recent work studies LLM-based juries for e-commerce disputes (Sun et al., 2026), decentralized prediction-market arbitration (Wen et al., 2026), and online dispute resolution (Tan et al., 2024), while arbitration institutions and blockchain protocols have begun deploying LLM adjudicators in practice (American Arbitration Association, 2025; GenLayer, 2026). The appeal is straightforward:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

many disputes require semantic judgment, such as whether an item matches its description, yet are too low-value to justify human review.

However, adjudication differs from other uses of LLMs in one way: the judge is chosen before the disputes it will rule on occur. For example, a marketplace might commit to a dispute-resolution model, or a contract might name an LLM in its arbitration clause. At that moment, the only evidence about a judge is its record. The choice of model matters: Ast et al. (2026) ran the same 99 consumer disputes through two frontier models and found that one ruled for the consumer in 3 cases and the other in 14. In other words, selecting an arbitrator has become a model-selection problem.

The standard answer to this problem is *reference agreement*: how often the judge's rulings match those of the institution it would replace. We argue that reference agreement is not sufficient to make this choice. For instance, consider two candidate judges, J_A and J_B, with the same reference agreement. After reading the same evidence, a prospective party can accurately forecast J_A's rulings, while J_B remains difficult to anticipate. If the anticipated behavior of J_A is unacceptable, the party can reject it, renegotiate the agreement, or demand a larger margin before consenting. None of these options is available for J_B, which makes commitment risky. In other words, predictability lets a party understand its risk and commit to an agreement.

This distinction is older than machine learning. Legal realism identifies the law, from the perspective of the affected party, with the prediction of what courts will do in fact (Holmes, 1897). Similarly, Fuller (1964) counts congruence between declared rules and official action among the minimal conditions for a system of rules to function as law at all. Under this reading, **a judge whose rulings cannot be anticipated is unsuitable for binding adjudication**, regardless of how often it agrees with a reference. We thus ask:

> *Are LLM adjudicators of comparable accuracy equally predictable to the parties they bind?*

To answer this question, we introduce **adjudicative legibility**: how well an external observer can predict a judge's ruling on an unseen dispute, given the dispute and, optionally, the judge's published rules and a finite record of its past rulings. Legibility requires no ground truth, since the target is predicting the behavior of the judge and not determining the correct outcome.

We study the adjudicative legibility of eight LLM adjudicators on contested real-world cases involving commercial disputes (Sun et al., 2026) and Wikipedia deletion debates.[1] We find that on more than a third of these disputes, the outcome depends on which adjudicator is chosen: two adjudicators rule for opposite parties in every one of five reruns. Even adjudicators closely matched on reference agreement rule differently on up to a third of disputes (Figure 1). Adjudicators of comparable accuracy also differ by up to $2.35\times$ in illegibility, a gap that survives simultaneous confidence intervals for at least one observer in each institution. The safety margin our bound prescribes against their illegibility, a third to half of the maximum stakes, is comparable to the margin with no forecast at all.

We also find that the standard remedies for opacity do not close the gap: short procedural rules perturb observers rather than inform them, past rulings teach a judge's base rate but not what distinguishes it from other judges, and observers read a consistent policy even into a judge that decides at random.

Key contribution. We show that reference agreement is not a sufficient property for LLM adjudicators, and introduce adjudicative legibility as the missing criterion: judges of comparable accuracy differ substantially in how well a party can anticipate them.

Concretely, our contributions are:

- **Legibility as a criterion for machine judges.** We define legibility as how well an observer can forecast a judge's rulings, and prove that a party can protect itself by demanding a safety margin that grows with the judge's illegibility, the *price of illegibility* (Section 2).
- **An accuracy-legibility map of LLM adjudicators.** Across two institutions, agreement does not imply legibility, and legibility depends on the observer (Section 4).
- **An analysis of disclosure.** Neither short procedural rules nor records of up to 16 rulings make an adjudicator legible; whether substantive rules do remains open (Section 4).

[1] Our source code is available at `https://anonymous.4open.science/r/adjudicative-legibility/README.md`.

2 ADJUDICATIVE LEGIBILITY

2.1 DISPUTES, PRESENTATIONS, AND ADJUDICATORS

Let $x \sim \mathcal{D}$ denote the *semantic content* of a dispute: the applicable rules, the parties' claims, and the evidence record. Let $t \sim \nu$ denote a *framing* (the order in which the parties appear, the ordering of evidence, phrasing and formatting choices), and let $z = t(x)$ be the resulting *case record*. Rulings take values in $\{1, 0\}$, corresponding to either party winning the case.[2]

An *adjudicator* J is a frozen protocol for determining the winning party: a model, its weights, the constitution it rules under, and its decoding parameters. On a record z, it rules $Y_J(z) \sim \text{Bernoulli}(p_J(z))$, where $p_J(z)$ is the probability of ruling 1. The adjudicator rules on each case in isolation and carries no memory between cases. Where the source institution reached a decision, we write $y^\star(x) \in \{1, 0\}$ for its verdict, and $y^\star(x) = \perp$ where it did not (e.g., a discussion closed without consensus). *Reference agreement* is $\mathcal{A}(J) = \Pr[Y_J(z) = y^\star(x) \mid y^\star(x) \neq \perp]$. Note that $\mathcal{A}$ measures conformity to the source institution, not correctness.

From the perspective of a party, a frozen adjudicator is unpredictable for three distinct reasons:

- (D) **Decoding.** There might be inherent randomness in the decision process, such as due to model temperature. In other words, $p_J(z)$ is not strictly 0 or 1.
- (P) **Presentation.** $p_J(z)$ can vary with the framing t of the same content x (e.g., due to different ways of structuring an argument). A party at contract time does not know how a dispute will be framed; we do not estimate this source (Section 4).
- (Π) **Policy.** $p_J(z)$ can differ from what an observer expects of a competent adjudicator reading z under the same rules.

The first two are properties of the adjudicator alone. The third, which is the focus of this paper, is a relation between the adjudicator and an observer.

2.2 OBSERVERS AND LEGIBILITY

An *observer* is a pair (I, F), where I is an interface that fixes what the observer is shown, and F is a forecaster that turns what it is shown into a probability. Given a new record z, the interface may additionally disclose two kinds of information about the adjudicator: its published *rules*, i.e., the constitution it rules under, and a *record* of size $B \geq 0$, $H_B = \{(z_b, y_b, \rho_b)\}_{b=1}^{B}$, where y_b is the adjudicator's ruling on case z_b and ρ_b is the public rationale it issued with that ruling. The observer outputs a forecast $q_{F,J,B}(z) \in [0, 1]$ of the probability that J rules 1. We describe the different observers and conditions we test in Section 3.

Definition 1 (Adjudicative legibility)**.** *The* legibility *of adjudicator J to observer (I, F) at budget $B \geq 0$ is the Brier score (Brier, 1950) of the observer's forecast against the adjudicator's fresh ruling,*

$$L_{F,J}(B) \;=\; \mathbb{E}\big[(q_{F,J,B}(z) - Y_J(z))^2\big],$$

where the expectation is over the dispute, the record, the adjudicator's ruling, and any randomness in F.

In other words, legibility is how well a party's agent can anticipate the adjudicator from what it is shown (lower is more legible). The definition makes no reference to $y^\star$: the target is what this adjudicator will do, not the correct outcome. It is relative to an observer, which we view as a feature, since an adjudicator that is legible to a well-resourced platform may be opaque to a consumer.

Part of the observer's error is independent of the choice of observer. If the adjudicator itself is random on a case, no forecast can do better than $p_J(z)$, and the Brier score separates the two:

Proposition 1 (Decomposition)**.** *For any observer whose forecast on a held-out case z is independent of $Y_J(z)$ given z and what the observer is shown,*

$$L_{F,J}(B) \;=\; \underbrace{\mathbb{E}\big[(q_{F,J,B}(z) - p_J(z))^2\big]}_{\varepsilon_{F,J}(B)\text{: illegibility}} \;+\; \underbrace{\mathbb{E}\big[p_J(z)\,(1 - p_J(z))\big]}_{\kappa_J\text{: own noise}}.$$

[2] Binary rulings reflect our two settings and simplify our analysis, which also applies in richer outcome spaces.

Only the first term depends on the observer or on what it is shown.

We call $\varepsilon_{F,J}(B)$ the *illegibility* of J to the observer: it is the observer's error about the adjudicator's policy, net of the adjudicator's own randomness, and it is the only part of L that disclosure can reduce. Consequently, the gain from any disclosure (e.g., $L_{F,J}(0) - L_{F,J}(B) = \varepsilon_{F,J}(0) - \varepsilon_{F,J}(B)$) does not depend on κ_J: when we compare two disclosure conditions, the adjudicator's own noise is the same in both and cancels out.

2.3 Why legibility matters: the price of illegibility

We conclude this section by showing the value of legibility for a party deciding whether to accept J as the adjudicator for its future disputes. Let $u_i(y, z)$ be the party's utility when the ruling on z is y, let $a_i(z) = u_i(1, z) - u_i(0, z)$ be the stakes, and suppose $|a_i(z)| \le S$. The party's expected value under J is $V_i(J) = \mathbb{E}[u_i(0, z) + a_i(z)\, p_J(z)]$. The party cannot compute this, since it does not know p_J. What it can compute, from what it is shown, is the same quantity with its forecast in place of the adjudicator, $\widehat{V}_i(J) = \mathbb{E}[u_i(0, z) + a_i(z)\, q_{F,J,B}(z)]$.

Proposition 2 (Price of illegibility)**.** *Under the assumptions of Proposition 1,* $\left|\widehat{V}_i(J) - V_i(J)\right| \le S\sqrt{\varepsilon_{F,J}(B)}$*. Hence, if* d_i *is the party's outside option and* $r \ge \varepsilon_{F,J}(B)$*, then* $\widehat{V}_i(J) \ge d_i + S\sqrt{r}$ *implies* $V_i(J) \ge d_i$*.*

In other words, a party can protect itself against an adjudicator it does not fully understand by demanding a margin equal to the stakes times the square root of the adjudicator's illegibility, $S\sqrt{\varepsilon_{F,J}(B)}$. Note that only ε enters the bound: the adjudicator's own noise adds variance to a party's outcomes, but it does not mislead the party about what it is agreeing to, whereas illegibility does. Note also that the bound holds in expectation over $\mathcal{D}$ and not for every dispute, so r must be an upper confidence limit rather than a point estimate. Proofs are in Appendix A.

3 Experimental Design

We design our experiments to answer four research questions (full setup in Appendix B):

- **RQ1 Does the choice between adjudicators of similar accuracy change the outcome of disputes?** *Rationale:* if so, agreement hides consequences of the choice.
- **RQ2 Can a party anticipate the behavior of adjudicators of similar accuracy equally well?** *Rationale:* our central claim is that legibility varies where accuracy does not.
- **RQ3 Does legibility depend on the observer?** *Rationale:* an adjudicator's position on the map may reflect the observer rather than the adjudicator.
- **RQ4 What makes an adjudicator more or less legible?** *Rationale:* parties and institutions can control what is disclosed; we test whether rules and past rulings help.

3.1 Adjudication Settings

We chose two settings that differ in language, evidence, and how the source institution decides, so that a property shared across both is unlikely to be an artifact of either.

VerdictBench (Sun et al., 2026) is made of Chinese-language e-commerce disputes between a buyer (Party A) and a seller (Party B), each decided by a panel of 17 crowd jurors. A case record contains the admissible text and every attached image; we exclude cases with video. The reference $y^\star$ is the juror majority.

Articles for Deletion (AfD) is the English Wikipedia process for deciding whether an article is deleted (Party A) or kept (Party B). Editors post policy-based arguments, and an uninvolved closer assesses rough consensus (not through voting). We render each discussion as a card: each side's substantive arguments in chronological order, stripped of identities, signatures, timestamps, and closure language. The reference $y^\star$ is the closer's decision, and $y^\star = \perp$ where the closer found no consensus.

Sampling on contested disputes. In both settings, we sample disputes on which the source institution was itself divided: VerdictBench cases by juror margin (at most 7 of 17) and AfD

discussions by the depth of opposition, i.e., the number of substantive positions on the weaker side. This is a deliberate choice of estimand: on easy disputes competent adjudicators rule alike, so the choice matters most on contested ones, which are often the source of risk in agreements. From each contested pool, we draw disjoint sets of TEST cases, on which adjudicators are forecast, HISTORY cases, from which records are sampled, and CALIBRATION cases, used for every adaptive decision (prompts, parsers, control construction). We evaluate on 100 TEST cases per dataset. On AfD, 21 of these closed without consensus: they are still used to compute legibility, since the target is the adjudicator's ruling, but not to compute reference agreement. The total cost of our experiments, between compute and API use, was $\sim$5.9k USD.

3.2 ADJUDICATORS

We evaluate eight adjudicators, J1–J8, from eight model families, spanning frontier and mid-tier models: GPT-5.6 Sol, Claude Opus 5, Gemini 3.1 Pro, Mistral Medium 3.5, Kimi K3, Grok 4.6, Muse Spark 1.2, and Qwen3.8 27B. Each adjudicator is a frozen protocol, with a pinned model snapshot and provider route, a short dataset-specific constitution, and a fixed decoding policy. For each case, an adjudicator reads the complete record (text and, for VerdictBench, all images), issues a short public rationale, and rules for A or B. The adjudicator has no memory across cases, no tools, and no access to the reference. We rerun every adjudicator five times on each TEST case to estimate $p_J(z)$ and its own noise κ_J; HISTORY cases receive a single ruling, since a party observes the record as it was actually decided.

Controls. Two constructed adjudicators test alternative explanations (Appendix E): C1, which rules by a seeded coin, tests whether observers can recognize an adjudicator with no policy, and C2, J7 under a deliberately vague constitution with matched agreement, tests whether legibility merely tracks how specific the rules are.

3.3 OBSERVERS

Our primary observers are LLM agents. For each target case, a fresh agent receives the complete target record, the adjudicator's constitution (if the setting allows it), and a set of tools to explore exactly B past cases of one adjudicator: it can list them, search them, and open any one of them together with the ruling and rationale the adjudicator issued. It has no web access, no memory across targets, and no knowledge of which model it is forecasting. It ends with a probability that the adjudicator rules for Party A.

We use two agents from different model families, F1 (GPT-5.6 Luna) and F2 (Gemini 3.7 Flash), so that no finding depends on a single observer; no observer is the same model as any adjudicator. Additionally, two deterministic observers serve as baselines. F3 is a logistic model over frozen text and image embeddings (Chen et al., 2024; Tschannen et al., 2025) of the past cases and their rulings, and measures what a simple representation can recover from the same record. F4 predicts the adjudicator's smoothed base rate from the B rulings it is shown and ignores the case entirely: it is the baseline against which any case-specific learning is measured.

Disclosure conditions. We compare three conditions: (i) *dispute only*, in which the observer sees the target case and nothing else; (ii) *rules*, in which it additionally sees the adjudicator's constitution, exactly as given to the adjudicator; and (iii) *rules and record*, in which it additionally sees $B \in \{4, 16\}$ past rulings with their rationales, drawn from HISTORY under two independent seeds. Since all eight adjudicators share one constitution per dataset, conditions (i) and (ii) produce a single forecast per case, which we score against every adjudicator. For the same reason, every forecast in condition (iii) can be rescored against the rulings of every other adjudicator on the same cases, which tells us whether a record teaches the individual adjudicator or only the task.

3.4 MEASUREMENT

Illegibility (Proposition 1) compares the observer's forecast with the adjudicator's ruling probability $p_J(z)$, which we cannot observe directly. We estimate it from the five reruns of each test case, as the fraction $\hat{p}$ of rulings for Party A. Since five reruns make $\hat{p}$ look more decisive than $p_J(z)$, and would thus overstate illegibility, we apply the standard finite-sample correction (Ferro, 2014), which makes

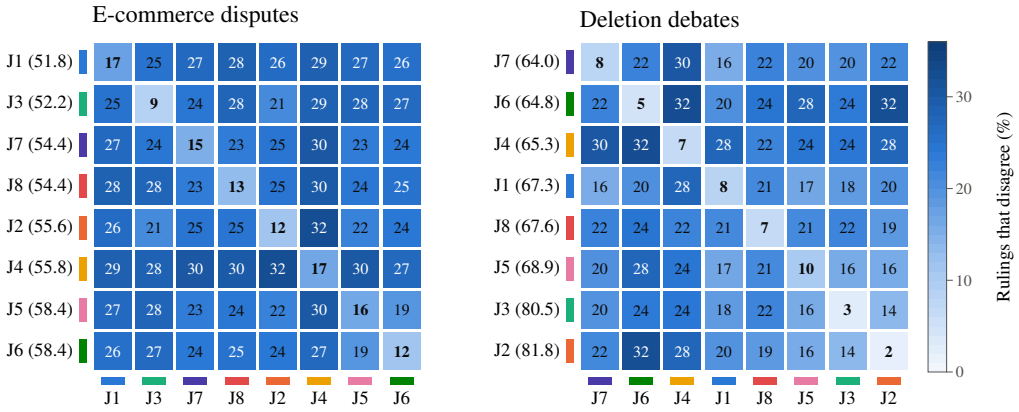

Figure 2: **Adjudicators disagree where agreement does not.** Probability (%) that two independent rulings from the row and column adjudicators disagree. Rows are ordered by reference agreement (in parentheses); the diagonal is each adjudicator's own noise.

both estimates unbiased (Appendix C). Reference agreement is the expected agreement between $\hat{p}$ and $y^\star$ over the cases where $y^\star$ is defined. Intervals are 95% percentile intervals from 1,000 paired bootstrap replicates over test cases and history seeds; comparisons between adjudicators additionally use simultaneous intervals, widened so that they hold jointly for all 28 pairs. To locate where a change in illegibility comes from, we split it into a *base rate* component, due to the observer's average forecast, and a *case-specific* component, due to how its forecasts vary across cases; refer to Appendix C for details.

4 Results and Discussion

RQ1 (Adjudicator choice changes outcomes, and agreement does not show it). On 36% of contested e-commerce disputes and 40% of contested deletion debates, one adjudicator rules for one party in all five reruns while another rules for the other party in all five. Randomness does not explain this: if the eight adjudicators ruled for Party A with the same probability on each case, differing only in decoding noise, it would happen on fewer than 1% of cases, and eight five-member juries simulated from each e-commerce case's juror votes do so on 3%. In other words, on more than a third of the disputes we study, the outcome depends on which adjudicator is chosen.

Reference agreement does not reveal these differences. On e-commerce disputes, all eight adjudicators agree with the jury majority on 51.8% to 58.4% of cases (Figure 2). J2 and J4 differ by only 0.2 points in agreement, yet two independent rulings from them disagree on 32% of cases [26%, 39%]; on deletion debates, J1 and J8 differ by 0.25 points and disagree on 21% [15%, 28%].

> **Key observation.** On more than a third of contested disputes, the choice of adjudicator decides the outcome, and adjudicators with nearly identical reference agreement disagree on up to a third of cases.

RQ2 (Equally accurate, unequally legible). Adjudicators closely matched on agreement differ substantially in how well a party can anticipate them (Figure 3; all values in Table 3). With the dispute alone, F1's illegibility on e-commerce disputes ranges from 0.074 for J5 to 0.174 for J3, a 2.35× ratio (2.68× when four runs of F1 are averaged into one forecast), for adjudicators within seven points of agreement. The other observer and dataset combinations show ratios between 1.54× and 1.75×. Because the ratio of the most to the least legible of eight adjudicators exceeds one even when all are equally legible, we test pairs instead: with intervals that hold jointly for all 28 pairs, F1 separates three pairs in each institution (on e-commerce disputes, all involving J3), and F2 one pair on deletion debates. The same test separates no pair by agreement on e-commerce disputes. On deletion debates, where agreement ranges from 64% to 82%, it separates ten, including two of F1's three; J5 and J6 differ in legibility but not measurably in agreement.

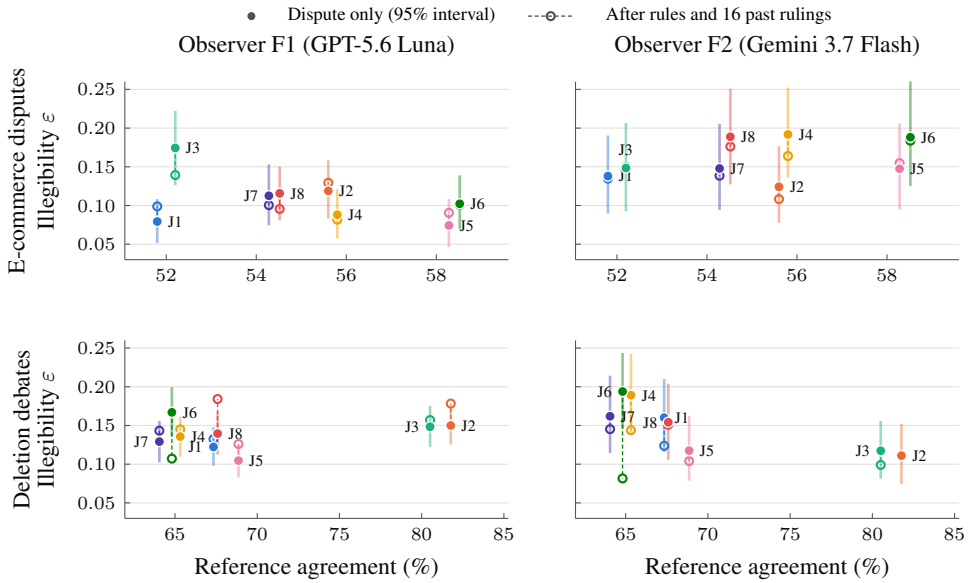

Figure 3: **Accuracy-legibility map across institutions and observers.** Filled: dispute only (95% intervals); open: after the rules and 16 past rulings. The observers rank adjudicators differently, and after 16 rulings the spread is still $1.7\times$ to $1.8\times$ in every panel (point estimates). Tied points are offset for display.

Note that, with the dispute alone, illegibility measures how far an adjudicator's policy departs from what the observer expects (source Π, Section 2.1); since records of up to 16 rulings teach no adjudicator-specific policy (RQ4), we do not claim that adjudicators differ in how learnable they are. Still, the gap is not an artifact of withholding the record, of randomness, or of calibration. After an adjudicator's rules and 16 of its rulings, the spread is still $1.7\times$ to $1.8\times$ in every panel of Figure 3. The adjudicator's own randomness is removed by construction (Proposition 1), and the observer's, since one forecast per case is scored against all eight adjudicators, can only shrink the ratio. Finally, recalibrating F1's forecasts without regard to which adjudicator is forecast leaves the ratio at $2.20\times$ (Appendix B.5).

By Proposition 2, illegibility has a price. At upper confidence bounds, as the proposition requires, the margin that protects a party ranges from 0.33 to 0.51 of the maximum stakes; a forecast of 0.5 on every case, which needs no observer, requires at most 0.5. In other words, on contested disputes, our observers' forecasts buy a party little protection.

Key observation. Adjudicators of similar accuracy differ by up to $2.35\times$ in illegibility, and legibility separates pairs in both institutions. On contested disputes, forecasting the outcome buys a party little protection.

RQ3 (Legibility depends on who is reading). We find no observer-independent ranking of adjudicators. With the dispute alone, the rank correlation between F1's and F2's illegibility is 0.00 on deletion debates and 0.05 on e-commerce disputes (no condition yields an interval excluding zero, although with eight adjudicators these estimates are imprecise). J2 and J4, 0.2 points apart in agreement, illustrate this: F1 anticipates J4 better, and F2 anticipates J2 better. Part of the difference lies in how the observers forecast. F1 hedges, with forecasts 28 points from 0.5 on average against 41 for F2, so it errs most on decisive adjudicators: across adjudicators, F1's illegibility falls as their own noise rises ($r = -0.87$ and -0.83), whereas F2's does not ($r = 0.03$ and 0.21).

Key observation. Our two observers rank adjudicators almost unrelatedly, partly because they forecast differently. Legibility is thus strongly linked to its observer.

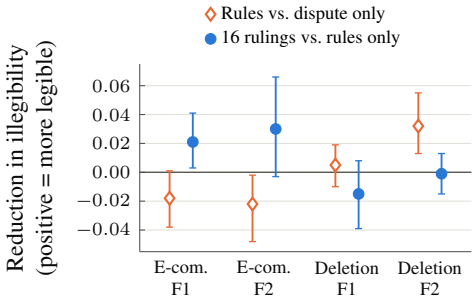

Figure 4: **Effect of disclosure.** Reduction in illegibility (positive: more legible) from adding the rules to the dispute (orange diamonds) and 16 past rulings to the rules (blue dots), averaged over J1–J8, with 95% intervals. Only one of the four effects of past rulings is clearly above zero, and the effect of the rules has opposite signs in the two institutions.

RQ4 (Small disclosures do not buy legibility). Neither disclosure a party can ask for, namely published rules and past rulings, closes the legibility gap (Figure 4).

Published rules. Our constitutions are short and procedural: roughly a hundred words asking for impartiality, attention to the evidence, and a binary ruling. Disclosing them nonetheless moves F2's forecasts by 7 to 8 points per case.[3] This reduces F2's illegibility by 0.032 [0.013, 0.055] on deletion debates but increases it by 0.022 [0.002, 0.048] on e-commerce disputes, and about 90% of either change is case-specific rather than a shift in base rate; F1 shows the same signs without resolved intervals. In other words, rules that do not decide cases do not inform the observer about the adjudicator: they perturb it, for better or worse. Since additional information cannot harm an ideal decision-maker (Blackwell, 1953), the decrease in forecasting performance is due to how our observers use the rules, not the rules themselves.

Past rulings. Providing sixteen past rulings with their rationales improves illegibility in only one of eight dataset, observer, and budget combinations (F1 on e-commerce disputes, 0.021 [0.003, 0.041]), and worsens it in one (F1 on deletion debates at $B = 4$). Moreover, what a record teaches is not specific to its adjudicator. We check this by scoring forecasts made from one adjudicator's record against the rulings of the other seven: if a record revealed its adjudicator's individual policy, these forecasts would fit that adjudicator better than the others. Beyond the advantage of knowing its base rate, which F4 also captures, they do not, for any adjudicator, observer, or dataset. The record teaches how adjudicators in general read the rules, much as demonstrations can help language models by conveying the label distribution and format rather than the input-output mapping (Min et al., 2022).

Controls. Observers also read a non-existent policy into noise: given C1, which publishes bare rulings drawn from a seeded coin, both forecast as though it followed a policy: after 16 rulings, their illegibility exceeds that of F4, which learns only the coin's rate, by 0.05 to 0.16. C2, J7 under a much vaguer constitution, is not measurably less legible than J7 (Appendix E).

Key observation. Neither short procedural rules nor records of up to 16 rulings make an adjudicator legible to our observers: the rules perturb them, and the records teach an adjudicator's base rate but not what distinguishes it from other adjudicators.

Limitations. Our core results rest on two LLM observers that behave differently (F2 is deterministic, while F1's forecasts vary between runs), and on a small design: 100 test cases per institution, records of at most 16 rulings, and two history seeds, so longer records may teach more. The rules we disclose are procedural, and whether substantive rules make adjudicators legible remains untested. We assume a party sees an adjudicator's record but cannot query it or learn its model; we therefore measure what disclosed behavior reveals, not what querying or reputation adds. F1 and J1, and F2 and J3, share model families, and Party A is always the buyer or the deletion side. We do not estimate sensitivity to presentation, and our results describe contested disputes from two institutions and specific model versions. Appendix F discusses each limitation in detail.

[3] F2 returned identical forecasts across repeated runs in our stability audit, so these shifts are not sampling noise (the audit used a record of 16 rulings).

5 Related Work

Predictability of adjudication. The predictability of adjudication predates LLMs: Holmes (1897) identifies the law, for the affected party, with the prediction of what courts will do, and Fuller (1964) requires congruence between declared rules and official action. Predictable rulings reduce transaction costs (Coase, 2013; North, 1991) and shape bargaining in their shadow (Mnookin & Kornhauser, 1979). In the terms of Kaplow (1992), our constitutions are standards rather than rules, which helps explain why disclosing them does not make an adjudicator predictable.

LLM adjudicators. LLMs already assist online mediation (Westermann et al., 2023; Tan et al., 2024), decide e-commerce disputes evaluated against crowd juries (Sun et al., 2026), resolve disputed prediction markets (Wen et al., 2026), and operate in deployed arbitration and contract-resolution systems (American Arbitration Association, 2025; GenLayer, 2026). All of these are evaluated by agreement with a reference. Ast et al. (2026) observe that models rule differently on the same consumer disputes and argue that arbitrator selection has become model selection. We provide the controlled measurement this calls for, and a criterion beyond agreement.

Evaluating and predicting judges. Work on LLM-as-a-judge asks whether a model's assessments are reliable, and has documented bias, instability, and manipulability (Zheng et al., 2023; Shi et al., 2025; Hu et al., 2026; Li et al., 2025); Thakur et al. (2025) show that agreement summaries can hide large differences between judges. Legal judgment prediction forecasts courts and individual justices from case text (Aletras et al., 2016; Katz et al., 2017). We ask instead whether a party can anticipate rulings, for separately developed adjudicators on identical disputes.

Multiplicity and simulatability. The fact that adjudicators of similar accuracy decide over a third of contested disputes in opposite ways is an instance of predictive multiplicity (Marx et al., 2020; D'Amour et al., 2022). Legibility adds an observer-dependent axis: disagreement measures whether two adjudicators differ, legibility whether a party can tell how. Predicting a model's behavior is an established paradigm (Doshi-Velez & Kim, 2017; Hase & Bansal, 2020); closest to us, ALMANACS finds that explanations do not consistently help predict an LLM's answers (Mills et al., 2025), and Shapira et al. (2026) find that target-adaptive learners predict unfamiliar agents better than prompted LLMs. We do not introduce simulatability: we apply it where a party must commit to an adjudicator in advance, contrast it with agreement, price forecast error as a contractual margin, and measure the value of disclosure.[4]

6 Conclusion

At the time of writing, machine adjudicators are chosen based on accuracy, and we showed that this property is not sufficient. On more than a third of contested disputes, the outcome depends on which LLM adjudicator is chosen, and adjudicators that agree almost equally often with the institution they replace differ by up to $2.35\times$ in illegibility, i.e., in how poorly a party can anticipate them. Common disclosures, such as procedural rules and small records, do not close the gap, and the safety margin our observers' forecasts justify is comparable to that with no forecast at all. We therefore argue for two changes. First, before an adjudicator is granted authority, it should be evaluated on both reference agreement and legibility, and the observer used to measure legibility should be reported. Second, legibility is unlikely to come from disclosing short procedural rules or a handful of past rulings: making an adjudicator legible will likely require rules specific enough to decide cases. We also stress that legibility is not fairness, and a legible adjudicator can be legibly biased; but only a legible one lets a party see the bias before it consents. In short, while accuracy tells a party how often an adjudicator agreed with the past, legibility tells it what it is agreeing to. As machine adjudicators take on more real-world applications, both properties will be key for robust and useful adjudicators.

[4]We use legibility only for whether rulings can be forecast; Appendix G discusses its other meanings in the literature (Dragan et al., 2013; Kirchner et al., 2024).

AI USE STATEMENT

In this work, we used generative AI tools to help develop the conceptual framework, formulate and write the proofs of our two propositions, refine our research questions, provide feedback on the experimental design, implement the experimental pipeline, clean and reformat the two datasets, and interpret results. We have not used generative AI tools to generate synthetic data, and translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to identify and summarize related literature, brainstorm framings and titles, suggest a structure for the paper, draft and edit parts of the text, format references, and create the figures and their plotting code. The tools we used were Claude (Fable 5.1 and Opus 5.5), ChatGPT (Sol 5.6 and Astra), and Codex (Sol 5.6 and Astra).

We have reviewed all AI-assisted work. The research question, the notion of adjudicative legibility, and the study design originated with the authors, who directed every stage of the work through repeated rounds of proposal, critique, and revision, and who rewrote and edited every section of the paper; no AI-generated text was included without review. Every number reported in the paper was checked against the outputs of the analysis pipeline, whose adjudicator and observer responses were audited and whose forecasts were independently replayed (Appendix B.5). The proofs were checked by hand, and every cited work was checked against its source. Note that LLMs are also the object of our study: the adjudicators J1–J8 and the observers F1 and F2 are experimental subjects, not research assistants, and their use is described in Section 3 and Appendix B. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix A contains complete proofs of both propositions. Appendix B specifies the datasets and their sampling, the adjudicator routes and snapshots, the constitutions, and the observer protocols; Appendix C specifies the estimators and bootstrap procedures, and Appendix B.5 the audits, including an independent replay of every forecast. The anonymous repository linked in Section 1 contains the source code and prompt templates.

REFERENCES

Stefano V. Albrecht and Peter Stone. Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems. *Artificial Intelligence*, 258:66–95, 2018. doi: 10.1016/j.artint.2018.01.002. URL https://www.cs.utexas.edu/~pstone/Papers/bib2html/b2hd-AIJ18-Albrecht.html.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93, 2016. doi: 10.7717/peerj-cs.93. URL https://peerj.com/articles/cs-93/.

American Arbitration Association. AAA-ICDR AI Arbitrator Now Available for Documents-Only Construction Cases. Press release, November 2025. URL https://www.adr.org/press-releases/aaa-icdr-ai-arbitrator-now-available/. Published November 3, 2025.

Dylan R. Ashley, Gael Le Lan, Changsheng Zhao, Naina Dhingra, Zhipeng Cai, Ernie Chang, Mingchen Zhuge, Yangyang Shi, Vikas Chandra, and Juergen Schmidhuber. Rpra: Predicting an llm-judge for efficient but performant inference, 2026. URL https://arxiv.org/abs/2604.12634.

Kevin D. Ashley. *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. MIT Press, Cambridge, MA, 1990. ISBN 9780262011143. URL https://www.sci.pitt.edu/people/kevin-ashley.

Kevin D. Ashley. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, 2017. doi: 10.1017/9781316761380. URL https://doi.org/10.1017/9781316761380.

Federico Ast, William George, and Robert Dean. The New Arbitrator-Selection Problem in the Age of AI: Choosing Which Model Decides Your Dispute. Kluwer Arbitration Blog, July 2026. URL https://legalblogs.wolterskluwer.com/arbitration-blog/the-new-arbitrator-selection-problem-in-the-age-of-ai-choosing-which-model-decides 22 July 2026.

Sher Badshah, Ali Emami, and Hassan Sajjad. Judge, Retrieve, or Abstain: Uncertainty-Guarded LLM Judging with Provable Risk Guarantees, 2026. URL https://arxiv.org/abs/2608.17994. Accepted at the Conference on Language Modeling (COLM), 2026.

Charles de Secondat baron de Montesquieu. *Montesquieu: The Spirit of the Laws*. Cambridge University Press, September 1989. ISBN 978-0-521-36974-9.

David Blackwell. Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Michael Broyde and Yiyang Mei. Don't kill the baby: The case for ai in arbitration, 2024. URL https://arxiv.org/abs/2408.11608.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Danielle Keats Citron. Technological Due Process. *Washington University Law Review*, 85(6):1249–1313, 2008. URL https://journals.library.wustl.edu/lawreview/article/id/6697/.

R. H. Coase. The Problem of Social Cost. *The Journal of Law & Economics*, 56(4):837–877, 2013. ISSN 0022-2186. doi: 10.1086/674872. URL https://www.jstor.org/stable/10.1086/674872.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://jmlr.org/papers/v23/20-1335.html.

Adrian de Valois-Franklin and Alex Bogdan. Rails: Verification-native clearing for agentic commerce, 2026. URL https://arxiv.org/abs/2606.08790.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning, 2017. URL https://arxiv.org/abs/1702.08608.

Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. Legibility and Predictability of Robot Motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 301–308, 2013. URL https://publications.ri.cmu.edu/legibility-and-predictability-of-robot-motion.

Kevin Du, Alexander Hoyle, Laura Ruis, and Acyr Locatelli. Legibility is Not Interpretability: Comparing Judged and Actual Importance in Chain-Of-Thought Reasoning, 2026. URL https://arxiv.org/abs/2609.04194.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140:1917–1923, 2014. doi: 10.1002/qj.2270. URL https://doi.org/10.1002/qj.2270.

Lon L. Fuller. *The Morality of Law*. Yale University Press, 1964.

GenLayer. GenLayer: The adjudication layer for the agentic economy. GenLayer documentation, 2026. URL https://docs.genlayer.com/.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Peter Hase and Mohit Bansal. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.491. URL https://aclanthology.org/2020.acl-main.491/.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4351–4367. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.390. URL https://aclanthology.org/2020.findings-emnlp.390/.

Oliver Wendell Holmes, Jr. The path of the law. *Harvard Law Review*, 10(8):457–478, 1897.

Yiran Hu, Zongyue Xue, Haitao Li, Siyuan Zheng, Qingjing Chen, Shaochun Wang, Xihan Zhang, Ning Zheng, Yun Liu, Qingyao Ai, Yiqun Liu, Charles L. A. Clarke, and Weixing Shen. LLMs on Trial: Evaluating Judicial Fairness for Large Language Models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/28ba12956bcffefac28a31b4a0ce82cb-Abstract-Conference.html.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UHPnqSTBPO.

Louis Kaplow. Rules Versus Standards: An Economic Analysis. *Duke Law Journal*, 42(3):557–629, 1992. URL https://scholarship.law.duke.edu/dlj/vol42/iss3/2/.

Daniel Martin Katz, Michael J. Bommarito, II, and Josh Blackman. A general approach for predicting the behavior of the Supreme Court of the United States. *PLOS ONE*, 12(4):e0174698, 2017. doi: 10.1371/journal.pone.0174698. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174698.

Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-Verifier Games improve legibility of LLM outputs, 2024. URL https://arxiv.org/abs/2407.13692.

Jon Kleinberg and Manish Raghavan. Algorithmic Monoculture and Social Welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118. URL https://doi.org/10.1073/pnas.2018340118.

Songze Li, Chuokun Xu, Jiaying Wang, Xueluan Gong, Chen Chen, Jirui Zhang, Jun Wang, Kwok-Yan Lam, and Shouling Ji. Llms cannot reliably judge (yet?): A comprehensive assessment on the robustness of llm-as-a-judge, 2025. URL https://arxiv.org/abs/2506.09443.

John Locke. *Second Treatise of Government*. Barnes & Noble Publishing, 2004. ISBN 978-0-7607-6095-6.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive Multiplicity in Classification. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6765–6774. PMLR, 2020. URL https://proceedings.mlr.press/v119/marx20a.html.

Edmund Mills, Shiye Su, Stuart Russell, and Scott Emmons. ALMANACS: A Simulatability Benchmark for Language Model Explainability, 2025. URL https://arxiv.org/abs/2312.12747v2. Version 2, revised 2 February 2025; first submitted in 2023.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Robert H. Mnookin and Lewis A. Kornhauser. Bargaining in the Shadow of the Law: The Case of Divorce. *Yale Law Journal*, 88(5):950–997, 1979. doi: 10.2307/795824. URL https://gretchen.law.nyu.edu/fac-articles/713/.

Douglass C. North. Institutions. *Journal of Economic Perspectives*, 5(1):97–112, March 1991. ISSN 0895-3309. doi: 10.1257/jep.5.1.97. URL https://www.aeaweb.org/articles?id=10.1257/jep.5.1.97.

Lorenzo Pacchiardi, Konstantinos Voudouris, Ben Slater, Fernando Martinez-Plumed, Jose Hernandez-Orallo, Lexin Zhou, and Wout Schellaert. Predictaboard: Benchmarking llm score predictability, 2025. URL https://arxiv.org/abs/2502.14445.

Yangda Peng, Yunjia Qi, Hao Peng, Haotian Xia, Guanzhong He, Xintong Shi, Richeng Xuan, Songyuanyi Lu, Yixian Liu, Zhichao Hu, Yuhong Liu, Lei Hou, Bin Xu, and Juanzi Li. Can llm-as-a-judge reliably verify rubrics in agentic scenarios?, 2026. URL https://arxiv.org/abs/2606.29920v1. Version 1, submitted 29 June 2026.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445315. URL https://doi.org/10.1145/3411764.3445315.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227. PMLR, 2018. URL https://proceedings.mlr.press/v80/rabinowitz18a.html.

Eilam Shapira, Moshe Tennenholtz, and Roi Reichart. Predicting decisions of ai agents from limited interaction through text-tabular modeling, 2026. URL https://arxiv.org/abs/2605.12411.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Yanhui Sun, Wu Liu, Haifeng Ming, Xinru Wang, Hantao Yao, and Yongdong Zhang. CyberJurors: A Multi-Agent Simulation Task for E-Commerce Disputes Verdict, May 2026. URL http://arxiv.org/abs/2605.28369.

Jinzhe Tan, Hannes Westermann, Nikhil Reddy Pottanigari, Jaromir Savelka, Sebastien Meeus, Mia Godet, and Karim Benyekhlef. Robots in the middle: Evaluating llms in dispute resolution. *arXiv preprint arXiv:2410.07053*, 2024.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pp. 404–430. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.gem-1.33/.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, 2025. URL https://arxiv.org/abs/2502.14786.

Max Weber. *General Economic History*. Allen & Unwin, 1927.

Junhao Wen, Juncen Zhou, and Junjie Huang. Can llms help decentralized dispute arbitration? a case study of uma-resolved markets on polymarket, 2026. URL https://arxiv.org/abs/2604.15674.

Hannes Westermann, Jaromir Savelka, and Karim Benyekhlef. Llmediator: Gpt-4 assisted online dispute resolution, 2023. URL https://arxiv.org/abs/2307.16732.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

Lexin Zhou, Pablo A. Moreno-Casares, Fernando Martinez-Plumed, John Burden, Ryan Burnell, Lucy Cheke, Cesar Ferri, Alexandru Marcoci, Behzad Mehrbakhsh, Yael Moros-Daval, Sean O hEigeartaigh, Danaja Rutar, Wout Schellaert, Konstantinos Voudouris, and Jose Hernandez-Orallo. Predictable artificial intelligence, 2025. URL https://arxiv.org/abs/2310.06167v3. Version 3, revised 7 January 2025; first submitted in 2023.

A PROOFS

Proposition 1. Fix a record z and everything the observer is shown. By assumption, the forecast $q = q_{F,J,B}(z)$ is independent of the fresh ruling $Y = Y_J(z)$ given z and what the observer is shown. Since the adjudicator has no memory across cases, its fresh ruling on z does not depend on the record either, so $\mathbb{E}[Y \mid z, q] = p_J(z)$. Writing $p = p_J(z)$,

$$\mathbb{E}\big[(q-Y)^2 \mid z, q\big] = (q-p)^2 + 2(q-p)\,\mathbb{E}[p-Y \mid z, q] + \mathbb{E}\big[(p-Y)^2 \mid z\big] = (q-p)^2 + p(1-p).$$

Taking expectations over the dispute, the record, and any randomness in F gives $L_{F,J}(B) = \varepsilon_{F,J}(B) + \kappa_J$. The second term does not involve q, and therefore does not depend on the observer or on what it is shown.

Proposition 2. Write $a = a_i(z)$, $q = q_{F,J,B}(z)$, and $p = p_J(z)$. By definition, $\widehat{V}_i(J) - V_i(J) = \mathbb{E}[a(q-p)]$. By the Cauchy–Schwarz inequality,

$$\big|\widehat{V}_i(J) - V_i(J)\big| \le \sqrt{\mathbb{E}[a^2]}\,\sqrt{\mathbb{E}[(q-p)^2]} \le S\sqrt{\varepsilon_{F,J}(B)}.$$

Hence, if $r \ge \varepsilon_{F,J}(B)$ and $\widehat{V}_i(J) \ge d_i + S\sqrt{r}$, then $V_i(J) \ge \widehat{V}_i(J) - S\sqrt{\varepsilon_{F,J}(B)} \ge d_i + S\sqrt{r} - S\sqrt{\varepsilon_{F,J}(B)} \ge d_i$.

Remarks. Three remarks clarify the scope of the result. First, both expectations are taken over the same dispute distribution $\mathcal{D}$: the bound protects a party whose future disputes resemble those on which legibility was measured. Second, it holds in expectation and not for every dispute. Third, r must bound the population illegibility, which is why we evaluate the margin at upper confidence bounds in Section 4. Since $\varepsilon \le L$ by Proposition 1, a party that can observe only realized rulings, and cannot rerun the adjudicator to estimate its own noise, may use $S\sqrt{L}$ as a conservative margin. Proposition 1 is the standard decomposition of expected squared loss into error about the conditional probability and irreducible variance (Gneiting & Raftery, 2007). The bound of Proposition 2 is conservative: it allows forecast errors to be as correlated with the stakes as possible. If a party's stakes are instead independent of the forecast error, the misestimate reduces to $\mathbb{E}[a_i]\,\mathbb{E}[q-p]$, the base-rate component, which records do teach. Finally, a party that cares about the variance of its outcomes and not only their expectation should also consider total legibility $L = \varepsilon + \kappa$, which Table 3 allows to compute; under L, the coin-flip control C1 is no longer among the most legible adjudicators.

B FULL EXPERIMENTAL SETUP

Table 1 summarizes the setup; the rest of this appendix gives the full protocol.

Table 1: Summary of the experimental setup.

	VerdictBench	**Articles for Deletion**
Dispute	Buyer vs. seller, e-commerce	Delete vs. keep, Wikipedia article
Language, evidence	Chinese; text and images	English; text only
Source institution	17-juror crowd panel	Uninvolved closer, rough consensus
Reference $y^\star$	Juror majority	Closer's decision; $\perp$ if no consensus
Contestation criterion	Juror margin ≤ 7 of 17	Depth of opposition
TEST cases	100	100 (79 with defined $y^\star$)
Adjudicators J1–J8	GPT-5.6 Sol, Claude Opus 5, Gemini 3.1 Pro, Mistral Medium 3.5, Kimi K3, Grok 4.6, Muse Spark 1.2, Qwen3.8 27B	
Controls	C1: seeded coin; C2: J7 with a vague constitution, agreement-matched	
Reruns	5 per TEST case; 1 per HISTORY case	
Observers	F1, F2: research agents (GPT-5.6 Luna, Gemini 3.7 Flash); F3: embedding probe; F4: base rate	
Disclosure	Dispute only; rules; rules and record ($B \in \{4, 16\}$, 2 seeds)	
Primary metric	Illegibility ε: excess Brier over the adjudicator's own noise	

B.1 DATASETS

VerdictBench. We use VerdictBench (Sun et al., 2026) at pinned revision `701b67c`, released under CC-BY-NC-4.0. Of 6,000 source rows, 5,829 have a complete 17-juror vote vector, and 3,828 of these contain no video. We freeze every image by content hash, reject missing or corrupt images,

and deduplicate cases with identical admissible text and ordered images. Votes, staff adjudications, court fields, and any other outcome-bearing fields never enter the case record. For each case, we compute the absolute juror margin and select a contested core of 1,000 cases by raising the margin threshold one juror at a time until at least 1,000 cases qualify, then sampling uniformly within it. The final threshold is a margin of at most 7 of 17 jurors, at which 1,312 deduplicated cases qualify; for duplicate cases, votes are averaged before computing the margin. Each case text is capped at 550 tokens, with an explicit truncation marker where needed; images are supplied separately. We split the core at random into TEST (300), HISTORY (500), and CALIBRATION (200) sets. Contamination probes use disjoint cases outside the core. We evaluate the first 100 cases of the randomly ordered TEST set, in both datasets.

Articles for Deletion. We collect English Wikipedia AfD discussions through pinned revision identifiers, using daily deletion logs from July 1, 2025 to August 10, 2026 (discussions nominated in this window may close later). Each case is a single discussion, and its record is the exact revision immediately before closure. We retain discussions closed as Keep, Delete, or no consensus, with substantive arguments on both sides, and exclude merges, redirects, withdrawals, speedy deletions, and multi-article outcomes. To prevent repeat nominations of the same article from receiving multiple chances, we keep at most one discussion per article. We then rank discussions by *depth of opposition*, the number of substantive positions on the weaker side, and sample uniformly within the final tied tier. The frozen construct contains 1,961 discussions (1,544 closed as Keep or Delete, 417 without consensus), split into TEST (300), HISTORY (500), and CALIBRATION (200), with the remainder held out. The model-facing card contains the nomination and each side's top-level arguments in chronological order, with identities, signatures, timestamps, templates, and closure language removed. All qualifying arguments on both sides are retained after sanitization and deduplication; their number and length are not equalized between sides. Cards longer than 8,192 tokens are rejected rather than truncated, and the 100 TEST cards contain 207 to 1,407 tokens. The nominated article and the contents of cited sources are not included, since deleted articles cannot be recovered reliably and including them only for kept articles would leak the outcome.

B.2 ADJUDICATORS

Table 2 lists the adjudicators. Every request is pinned to a single provider route with fallbacks disabled, and the provider and model reported by the router are checked on every response. All adjudicators use their provider's high reasoning setting, except J8, whose route offers only a stronger setting, which we use. Each adjudicator returns a short public rationale followed by a final line of the form `RULING: A` or `RULING: B`. The published constitutions, which are identical to the constitution component of the adjudicators' prompts, are reproduced below; the complete prompt templates are in the supplementary material. Each case also includes a case-local resolution instruction, which appears in every condition, including dispute-only. On VerdictBench it reads: *"Decide whether the buyer's requested remedy or the seller's defense is better supported under the platform transaction record."* On AfD it reads: *"A Wikipedia closing administrator decides the discussion by assessing rough consensus under Wikipedia's deletion policies. This is not a vote: the strength and policy relevance of the arguments matter more than their number. Rule for Party A if the article should be deleted; rule for Party B if it should be kept. Use only the frozen record below and do not look up the article, discussion, participants, or eventual result."*

Table 2: Adjudicator roster. The same snapshot and route are used on both datasets. Routing service provided by OpenRouter.

ID	Model	Route
J1	GPT-5.6 Sol	OpenAI
J2	Claude Opus 5	Anthropic
J3	Gemini 3.1 Pro (preview)	Google Vertex
J4	Mistral Medium 3.5	Mistral
J5	Kimi K3	Moonshot AI
J6	Grok 4.6	xAI
J7	Muse Spark 1.2	Meta
J8	Qwen3.8 27B	AkashML

VerdictBench constitution (93 words).

> You are an impartial adjudicator of a two-party Chinese e-commerce dispute drawn from VerdictBench. Decide only which party's position is better supported by the complete supplied text and images. Party A is the buyer and Party B is the seller. Treat the parties symmetrically; do not favor a party because of its name, order, verbosity, language, or formatting. Apply the stated resolution rule first, then weigh relevant evidence for reliability and consistency. Do not assume facts absent from the record and do not use outside sources. Your ruling must be A or B.

AfD constitution (101 words).

> You are an uninvolved closing administrator assessing a frozen English Wikipedia Articles for Deletion docket. Party A argues to Delete the article and Party B argues to Keep it. Determine which position has stronger support in the supplied policy-based arguments. This is not a vote: argument quality, evidence, and policy relevance matter more than the number of apparent recommendations. The nominated article and linked-source contents are not in the record; do not browse or assume them. Merge, redirect, draftification, no consensus, and other third outcomes are unavailable in this strict binary construct. Your ruling must be A (Delete) or B (Keep).

B.3 OBSERVERS

LLM observers. F1 is GPT-5.6 Luna and F2 is Gemini 3.7 Flash, each pinned to a single route. Each forecast starts a fresh agent that receives the complete target record (including all target images) and, when relevant, the published constitution. For $B > 0$, the agent receives tools scoped to exactly B past cases of one adjudicator: `list_history` pages through the cases with a short synopsis and the adjudicator's ruling; `search_history` runs a deterministic lexical search over those cases only; `open_dossier` returns a cached summary of one past case together with the ruling and the adjudicator's rationale; `open_raw_text` returns the exact admissible text of one past case; and, on VerdictBench only, `open_raw_media_page` returns up to eight of its images. The same observer model produces dossiers from a single past case and its ruling, and they never see the target, the reference, or the adjudicator's identity. The agent has no web access, no memory across targets, and no other tools. It ends with a short rationale and a final line `P(A) = 0--100`. The observer prompt templates are in the code repository.

Baselines. F3 is an L2-regularized logistic regression ($C = 1$, no tuning) over fixed embeddings of the past cases, labelled with the adjudicator's rulings. Text is embedded with BGE-M3 (Chen et al., 2024) and, for VerdictBench, images with SigLIP 2 Base (Tschannen et al., 2025), at pinned revisions; the two are normalized and concatenated with equal weight. F3 sees neither rationales nor dossiers. F4 ignores the case entirely and predicts the Laplace-smoothed rate of Party-A rulings among the B past rulings it is shown, $(n_A + 1)/(B + 2)$, or 0.5 when $B = 0$. F3 and F4 receive exactly the same records as the matched LLM forecasts.

B.4 SAMPLING OF RECORDS

For each dataset and budget, history sets are drawn from HISTORY under fixed seeds and shared across adjudicators and observers, so that every forecast can be rescored against every adjudicator on the same test cases. We draw sets at different budgets independently rather than nesting them. Each HISTORY case carries a single ruling from each adjudicator, since a party observes the record as it was actually decided, while each TEST case is ruled on five times.

B.5 AUDITS

Observer stability. We reran a sample of forecasts with identical inputs. F2 returned identical forecasts. F1's repeated forecasts differ by a median of 0.06 (95th percentile: 0.37 on VerdictBench, 0.36 on AfD), so we report F1 but do not interpret its per-case changes. F1 also hedges more: its forecasts lie on average 28 points from 0.5, against 41 for F2, in every condition.

Repeated dispute-only forecasts. To check that F1's randomness does not drive the headline ratio, we reran its dispute-only forecasts on e-commerce disputes three more times. The ratio between the least and the most legible adjudicator is $1.79\times$, $2.65\times$, and $3.12\times$ in the individual runs, and $2.38\times$ [1.75, 3.55] for illegibility averaged over runs, against $2.35\times$ in the original run. J3 is the least and J5 the most legible adjudicator in every run. Individual runs therefore vary substantially, but the averaged ratio and the identity of the extremes match the original run. Averaging F1's four runs into a single forecast before scoring removes most of its own randomness and gives $2.68\times$ [1.89, 4.36]; its rank correlation with F2 remains 0.10 [-0.62, 0.67]. After a single recalibration of F1's forecasts, fitted with cross-fitting on TEST and pooled across adjudicators, the ratio is $2.20\times$ [1.79, 3.67], and J1 and J3 remain separated under simultaneous intervals.

Contamination. A text-continuation probe on 50 cases per dataset found exact continuations of at least 25 tokens on VerdictBench for F2 (3 of 50) and J2 (2 of 50), and none on AfD. All matches were 25–27-token spans of party labels, evidence-submission boilerplate, and timestamps. Memorization of VerdictBench cases by F2 and J2 therefore cannot be excluded, although the matched spans are template text.

Images in past cases. On VerdictBench, neither observer successfully opened an image from a past case (F2's 45 image requests all returned integrity errors), so records were effectively text-only.

Replay. Every forecast, including the 1,200 rules-only forecasts and the 300 repeated F1 forecasts, was independently replayed from the stored responses and reconciled with the reported results; no score changed.

Recovered forecasts. 289 of 19,600 forecast slots (1.5%) were recovered after operational failures, through bounded retries, format repairs, token-limit recovery, and feedback on invalid tool calls. Replay confirms the retained outputs, but not that recovery was score-neutral.

Rules-only condition. This condition was added after the main analysis, to separate the effect of rules from that of records, which the main conditions disclose together.

C STATISTICAL METHODS

Illegibility and own noise. Let $\hat{p}$ be the fraction of Party-A rulings among an adjudicator's $k = 5$ rulings on a test case. Since $\mathbb{E}[\hat{p}(1-\hat{p})] = \frac{k-1}{k}\,p(1-p)$, the unbiased estimate of own noise is $\frac{k}{k-1}\hat{p}(1-\hat{p})$, and the unbiased per-case estimate of illegibility for a forecast q is

$$\hat{\varepsilon}(z) = (q-\hat{p})^2 - \frac{\hat{p}(1-\hat{p})}{k-1}.$$

Both are averaged over test cases and then over history seeds. We never clip negative estimates. For the coin-flip control C1, p is known analytically, so $\hat{\varepsilon} = (q-q_0)^2$ and own noise is $q_0(1-q_0)$, with $q_0 = 0.441$ on VerdictBench and $q_0 = 0.5$ on AfD.

Reference agreement. For a case with defined reference $y^\star$, expected agreement is $\hat{p}$ if $y^\star = 1$ and $1-\hat{p}$ if $y^\star = 0$; reference agreement averages this over cases with defined references. No-consensus AfD cases are never assigned to either side.

Outcome dependence and disagreement. A case has *opposing majorities* if at least one adjudicator has $\hat{p} > 0.5$ and another has $\hat{p} < 0.5$; with five rulings there are no ties. A case has *confident disagreement* if one adjudicator has $\hat{p} = 1$ and another $\hat{p} = 0$. To separate policy from own randomness, we compare both rates with a null in which, on each case, all eight adjudicators share the pooled ruling probability and their five rulings are resampled (exact expectations, confirmed with 10,000 simulated panels): the null rates are 34.1% and 28.8% for opposing majorities, and 0.9% and 0.8% for confident disagreement. On e-commerce disputes, we also draw eight panels of five jurors from each case's 17-juror vote share: confident disagreement then occurs on 2.9% [2.6%, 3.1%] of cases, against 36% for the adjudicators. This simulates panels from vote shares; it does not measure disagreement between repeated human panels. For a pair of adjudicators j, k, the probability that two

independent rulings disagree is estimated by the mean of $\hat{p}_j + \hat{p}_k - 2\hat{p}_j\hat{p}_k$, which we report in the main text. Figure 7 additionally reports a noise-corrected distance between their ruling probabilities, the mean of $(\hat{p}_j - \hat{p}_k)^2 - \hat{p}_j(1-\hat{p}_j)/4 - \hat{p}_k(1-\hat{p}_k)/4$, which is zero for two adjudicators with identical ruling probabilities regardless of their own noise.

Intervals. All intervals are 95% percentile intervals from 1,000 bootstrap replicates that resample test cases, and history seeds where applicable, with replacement. The same draws are used across adjudicators, observers, and conditions, so that all contrasts are paired. Since each case carries one independent draw of each observer's forecast, resampling cases also propagates the observer's own randomness into every interval. The eight adjudicators are a fixed roster and are not resampled. For comparisons between adjudicators, we additionally compute simultaneous intervals over the 28 pairwise differences within each dataset, observer, and condition, using the bootstrap distribution of the maximum absolute studentized difference. These do not control error across datasets, observers, or conditions. Ratios between the most and least legible adjudicators reselect both in every replicate; since such ratios exceed one by construction, we report them as descriptive and base inference on the simultaneous pairwise intervals. Differences in reference agreement use the same procedure; on deletion debates, we resample the 79 cases that closed with consensus. That a pair is not separated by agreement does not show that the two adjudicators are equally accurate, only that our test cases cannot tell them apart. Rank correlations use average ranks for ties and are recomputed in every replicate.

Base rate and case-specific components. For a fixed adjudicator and observer, the mean squared difference between forecasts and ruling probabilities decomposes exactly as

$$\overline{(q-\hat{p})^2} = \left(\bar{q} - \bar{\hat{p}}\right)^2 + \overline{\left((q-\bar{q}) - (\hat{p}-\bar{\hat{p}})\right)^2},$$

where bars denote averages over test cases. We call the first term the *base rate* component and the second the *case-specific* component. We compute each term's change between two conditions for every adjudicator and average over adjudicators. The rerun correction cancels in these paired differences.

What a record teaches the individual adjudicator. Let $G(D, T)$ be the illegibility of forecasts made from the record of adjudicator D (the donor) when scored against adjudicator T (the target), at a given budget. The *individual component* of T is $I(T) = \mathrm{mean}_{D \neq T}\, G(D, T) - G(T, T)$: positive values mean that T's own record predicts T better than other adjudicators' records do. Since a record also reveals its adjudicator's base rate, we compute the same quantity for F4 and report $I_F(T) - I_{\mathrm{F4}}(T)$, the part of the individual component that goes beyond base rates.

D Additional Results

Table 3: Reference agreement, own noise, and illegibility ε (lower is more legible) per adjudicator, under the dispute-only (D), rules-only (R), and rules plus 16 past rulings (R+16) conditions. Point estimates; 95% intervals and the $B = 4$ condition are in Figures 5 and 6.

	VerdictBench								Articles for Deletion							
	Agr.	Noise	F1			F2			Agr.	Noise	F1			F2		
			D	R	R+16	D	R	R+16			D	R	R+16	D	R	R+16
J1	.518	.087	.079	.091	.099	.138	.158	.135	.673	.041	.122	.105	.132	.160	.122	.124
J2	.556	.058	.119	.163	.129	.124	.153	.108	.818	.010	.150	.138	.178	.111	.093	.111
J3	.522	.044	.174	.180	.139	.149	.159	.150	.805	.014	.148	.157	.157	.117	.090	.099
J4	.558	.083	.088	.107	.082	.192	.199	.164	.653	.033	.136	.142	.145	.189	.164	.144
J5	.584	.082	.074	.107	.091	.148	.192	.155	.689	.052	.105	.094	.126	.118	.088	.104
J6	.584	.058	.102	.101	.101	.188	.217	.184	.648	.024	.167	.161	.107	.194	.154	.082
J7	.544	.077	.113	.128	.100	.148	.187	.139	.641	.040	.129	.122	.143	.162	.119	.145
J8	.544	.066	.116	.131	.096	.189	.186	.176	.676	.033	.140	.137	.184	.154	.123	.151
C1	.511	.247	.096	.104	.091	.182	.198	.187	.500	.250	.042	.061	.070	.107	.106	.124
C2	.522	.065	.129	.132	.125	.136	.177	.128	.656	.038	.127	.117	.129	.152	.153	.146

Illegibility. Figures 5 and 6 report illegibility with 95% intervals for every observer, target, and condition, including the $B = 4$ condition and the base-rate observer F4. F3's illegibility at $B = 16$ is within 0.002 of F4's in both datasets (0.189 versus 0.191 on VerdictBench, 0.216 versus 0.217 on AfD), so we omit it.

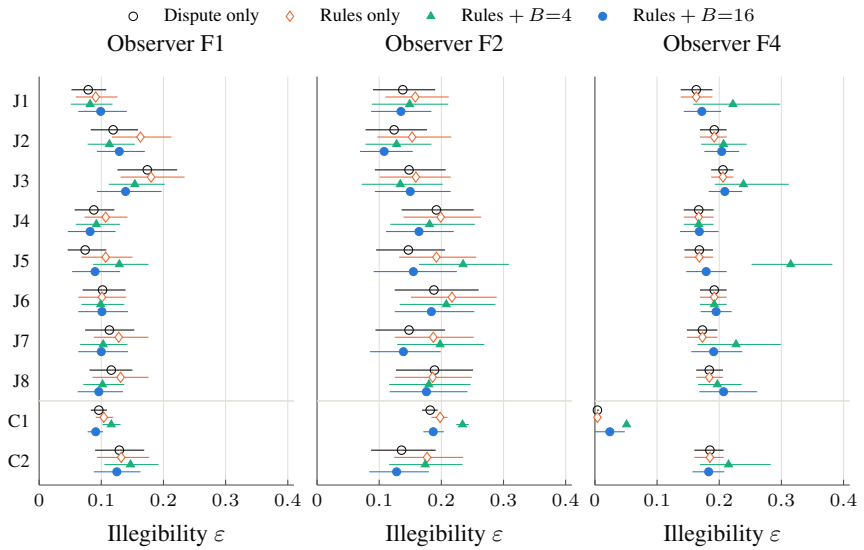

Figure 5: Illegibility ε with 95% intervals on e-commerce disputes, for every observer, target, and condition. F4 does not read rules, so its dispute-only and rules-only values coincide. Point estimates for the main conditions are in Table 3.

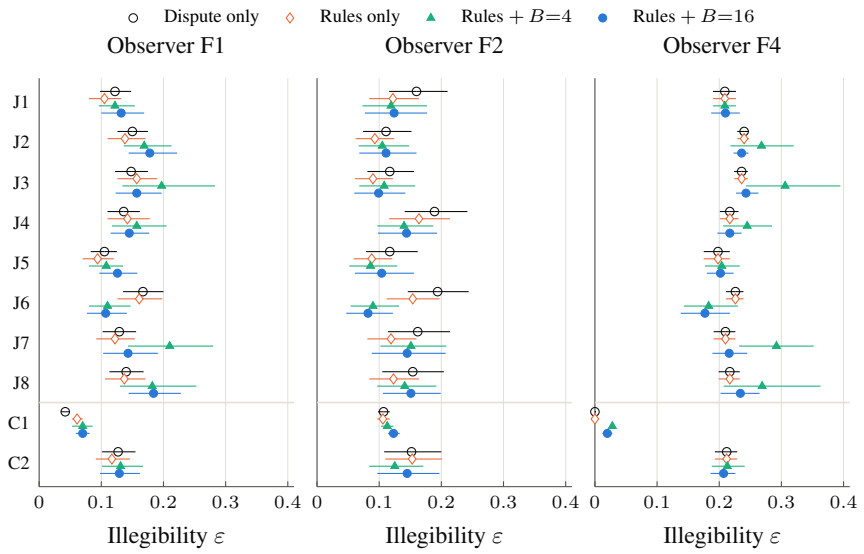

Figure 6: Illegibility ε with 95% intervals on deletion debates, for every observer, target, and condition. F4 does not read rules, so its dispute-only and rules-only values coincide.

Disagreement between adjudicators. Figure 7 reports the noise-corrected distance between adjudicators' ruling probabilities, and Figure 8 the pairs whose reference agreement differs by less than three points.

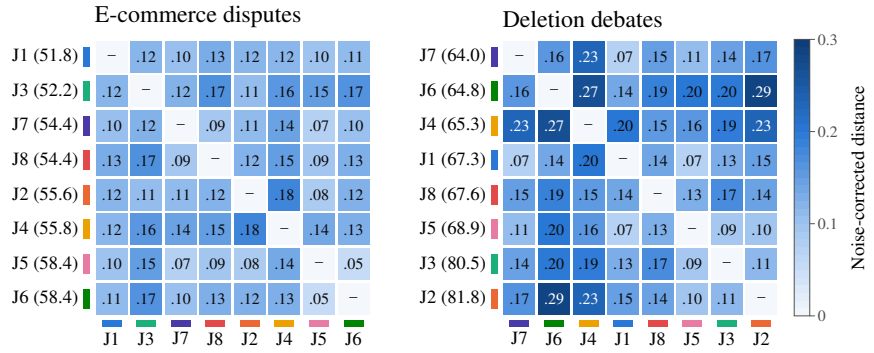

Figure 7: Noise-corrected distance between adjudicators' ruling probabilities (Appendix C); larger values mean the two adjudicators rule differently more often, and the distance is zero for adjudicators with identical ruling probabilities regardless of their own noise. Rows are ordered by reference agreement (in parentheses). 95% intervals are in the supplementary material.

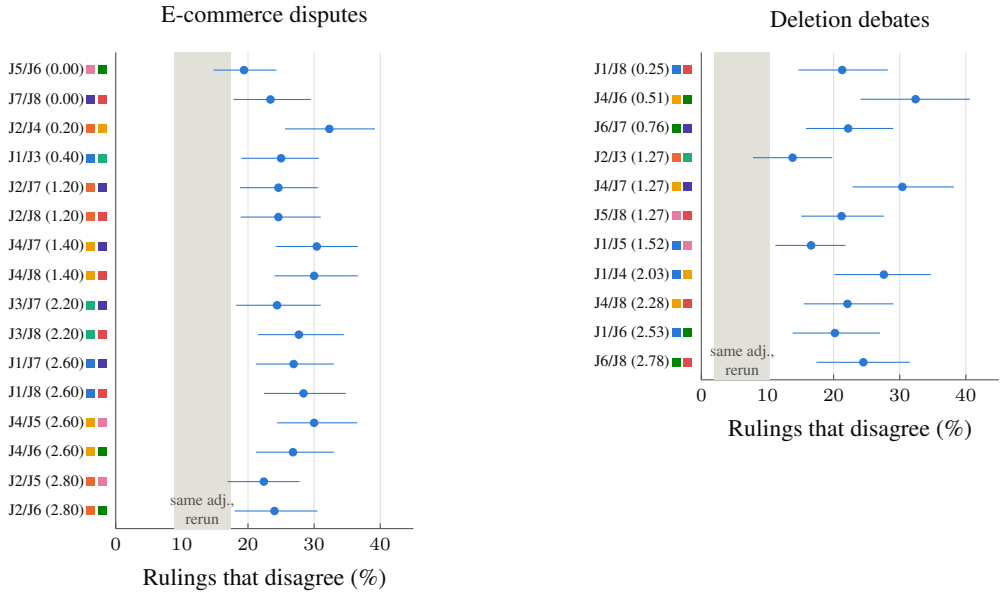

Figure 8: Adjudicator pairs whose reference agreement differs by less than three points (gap in points in parentheses), and the probability that two independent rulings from them disagree, with 95% intervals. The shaded band is the range of each adjudicator's disagreement with its own rerun. Pairs are selected on point estimates and are descriptive.

Rules and records. Table 4 reports the average effect of each disclosure over J1–J8. On deletion debates, the gain from disclosing the rules is almost entirely case-specific for F2 (0.028 of 0.032, interval $[0.012, 0.049]$), while its base rate component is not distinguishable from zero. Disclosing the rules changes F2's forecasts by 7.9 points per case on average on VerdictBench and 7.1 points on AfD. F1's changes are larger (12.7 and 10.7 points) but comparable to the variation between its own repeated forecasts (about 10 points), and the two observers' changes are uncorrelated across cases ($r = 0.00$ and 0.11).

Individual component. After subtracting F4's contribution, the individual component at $B = 16$ averages -0.001 (F1) and -0.001 (F2) on AfD, and $+0.006$ (F1) and -0.004 (F2) on VerdictBench. None of the 32 intervals for individual real adjudicators has a positive lower bound. Before subtracting F4, the component is often positive, but most of it is also present for F4, i.e., it reflects each adjudicator's base rate.

Table 4: Average reduction in illegibility from each disclosure, over J1–J8 (positive means more legible), with paired 95% intervals. Results involving the rules-only condition are from the follow-up described in Appendix B.5.

Dataset	Observer	Rules vs. dispute only	B=4 vs. rules only	B=16 vs. rules only
VerdictBench	F1	-0.018 [-0.038, 0.001]	0.017 [-0.001, 0.037]	0.021 [0.003, 0.041]
VerdictBench	F2	-0.022 [-0.048, -0.002]	0.005 [-0.023, 0.034]	0.030 [-0.003, 0.066]
AfD	F1	0.005 [-0.010, 0.019]	-0.025 [-0.055, -0.000]	-0.015 [-0.039, 0.008]
AfD	F2	0.032 [0.013, 0.055]	0.002 [-0.015, 0.017]	-0.001 [-0.015, 0.013]

Observer agreement. Table 5 reports the rank correlation between F1's and F2's illegibility across J1–J8.

Table 5: Spearman correlation between F1's and F2's rankings of J1–J8 by illegibility, with 95% intervals.

Dataset	Dispute only	Rules only	Rules + B=4	Rules + B=16
VerdictBench	0.05 [-0.67, 0.64]	-0.43 [-0.81, 0.45]	-0.10 [-0.55, 0.60]	-0.36 [-0.79, 0.52]
AfD	0.00 [-0.33, 0.74]	0.43 [-0.45, 0.76]	0.74 [-0.05, 0.88]	0.52 [-0.31, 0.86]

No-consensus cases. On AfD, F2's gain from dispute only to $B = 16$ is 0.021 on the 79 cases closed as Keep or Delete and 0.068 on the 21 cases closed without consensus, while F1's gains are -0.010 in both strata. We do not claim a statistically established difference between strata.

E CONTROLS

Our protocol included four constructed adjudicators, each differing from a real adjudicator in one declared respect. The main text uses C1 and C2; we also report C3 and C4, which did not work as intended.

C1: no policy. C1 rules for Party A with fixed probability q_0 on every case, independently of the case. It publishes only its sampled ruling, with no rationale: a fabricated justification would introduce apparent policy, and a statement that the ruling was random would reveal the control. Its records therefore contain case summaries and rulings but, unlike those of the real adjudicators, no rationales. An observer that learned only this rate would approach zero illegibility; F4 does so, with an expected excess of $[Bq_0(1-q_0) + (1-2q_0)^2]/(B+2)^2$ at budget B. The LLM observers do not: at $B = 16$, their illegibility exceeds F4's by 0.050 (F1) and 0.103 (F2) on AfD, and by 0.067 (F1) and 0.163 (F2) on VerdictBench. On VerdictBench, once each adjudicator's own randomness is removed, three to five of the eight real adjudicators are, in the point estimates, no more predictable to each observer than the coin.

C2: vague constitution. C2 is J7 under a 21-word constitution ("Decide fairly between Party A and Party B based only on the supplied record. Your ruling must be A or B."), tuned on CALIBRATION to match J7's reference agreement. The realized agreement gap on TEST is 1.5 points on AfD and 2.2 on VerdictBench. The change of constitution altered J7's behavior only modestly: their majority rulings differ on 9 and 10 of 100 cases, and the noise-corrected distance between their ruling probabilities is 0.024 [0.006, 0.046] on AfD and 0.007 [-0.005, 0.021] on VerdictBench. No difference in illegibility between C2 and J7 has an interval excluding zero.

C3: planted doctrine. C3 is J1 with a planted rule appended to its constitution: every case carries a synthetic marker, CEDAR or SLATE, and on CEDAR cases C3 must rule for Party A regardless of the merits. The rule has a clear effect: on the cases where it applies (49 of 100 on AfD and 62 of 100 on VerdictBench), C3's ruling probability differs from J1's by 0.36 and 0.52 on average. However, the marker was added only to the adjudicator's input, and was absent from every record shown to observers, both for the target case and for past cases. Observers could encounter the rule only indirectly, when C3's own rationale mentioned it: this happened in none of the episodes at

$B = 4$ and in about half of the episodes at $B = 16$ on VerdictBench. Since observers could never see whether the rule applied to the target, C3 tests neither recovery of a visible rule from the record nor use of a disclosed rule, and we do not use it. Testing whether substantive, checkable rules make adjudicators legible remains the most direct follow-up to our work.

C4: decoding noise. C4 is J4 decoded at temperature 1, intended to raise own noise without changing policy. The manipulation was too weak to pass its check on CALIBRATION: on TEST, C4's own noise exceeds J4's by 0.021 on AfD and falls short of it by 0.005 on VerdictBench. We therefore treat C4 as uninformative.

F EXTENDED LIMITATIONS

We stress that our results are subject to five main limitations. First, legibility is relative to an observer, and we study four. The two LLM-based observers differ in more than model family: F2 returns identical forecasts on repeated runs, whereas F1's forecasts vary by about ten points between runs (Appendix B.5), so we do not interpret F1's per-case changes. Other observers, such as humans or learners trained on an adjudicator's behavior (Shapira et al., 2026), may place adjudicators differently. Second, our design is small: 100 test cases per institution, records of at most 16 rulings, and two history seeds. Our finding that past rulings do not reveal what distinguishes one adjudicator from another therefore holds for records of this size, and longer records may teach more. Third, the rules-only condition was added after the main analysis; our reading of F2's shifts assumes that its determinism, observed with a record of 16 rulings, also holds without one. Moreover, the rules we disclose are short and procedural, as we expect deployed ones to be; whether substantive rules make adjudicators legible remains untested (Appendix E). Fourth, we do not estimate sensitivity to presentation (Section 2.1), which a party at contract time also faces, and on e-commerce disputes observers never opened the images of past cases, so the record they read was effectively text. Finally, our estimand is contested disputes from two institutions that differ in language, evidence, and decision process, and our results describe specific model versions. A contamination probe found short exact continuations of VerdictBench template text for F2 and J2, and some forecasts were recovered after operational failures (Appendix B.5).

G EXTENDED RELATED WORK

Predictability in legal and economic theory. The demand that adjudication be predictable predates machine learning by centuries. baron de Montesquieu (1989) and Locke (2004) argue for protection from arbitrary rulings, and Holmes (1897) identifies the law, for the affected party, with the prediction of what courts will do in fact. Fuller (1964) counts publicity and congruence between declared rules and official action among the conditions a system of rules must satisfy to function as law. In economics, predictable rules and rulings support rational economic organization (Weber, 1927) and reduce the transaction costs that parties incur (Coase, 2013; North, 1991), and expectations about adjudicated outcomes shape the agreements parties reach in its shadow (Mnookin & Kornhauser, 1979). Kaplow (1992) distinguishes rules, which fix the content of the law in advance, from standards, which leave it to be settled when the law is applied. Our constitutions are standards in this sense, which helps explain why disclosing them does not make an adjudicator predictable (Section 4). Predictability is not the same as legitimacy: Citron (2008) identifies the notice, participation, and review that automated decisions owe to the people they affect, none of which a successful forecast provides. We treat legibility as one measurable criterion for choosing an adjudicator, not as a complete account of fair adjudication.

LLMs in adjudication and dispute resolution. LLMs have been used to assist online mediation, reformulate party messages, and propose mediator interventions (Westermann et al., 2023; Tan et al., 2024), and legal scholarship has argued that consensual arbitration is a natural setting for AI adjudicators, since parties choose the procedure in advance (Broyde & Mei, 2024). Empirical work has moved closer to automated adjudication. CyberJurors introduces VerdictBench and evaluates multi-agent verdicts against crowdsourced juries (Sun et al., 2026), Wen et al. (2026) measure agreement between LLMs and UMA's resolutions of disputed prediction markets, and RAILS places semantic verification inside a clearing architecture for agentic commerce (de Valois-Franklin &

Bogdan, 2026). In practice, the AAA-ICDR's AI arbitrator operates under human review (American Arbitration Association, 2025), and GenLayer resolves contract outcomes by consensus among LLM validators (GenLayer, 2026). All of these evaluate adjudicators by agreement with a reference. Closest to our motivation, Ast et al. (2026) observe that models and model versions rule differently on the same consumer disputes and argue that arbitrator selection has become model selection. We provide the controlled measurement this argument calls for, and a criterion beyond agreement on which to make the choice.

Predicting judges. Legal judgment prediction forecasts court outcomes from case text (Aletras et al., 2016), and Katz et al. (2017) forecast the votes of individual Supreme Court justices, so the idea of predicting a particular decision-maker is not new. Case-based reasoning systems treat precedents as arguments about how a case should be decided (Ashley, 1990; 2017). In our setting, past rulings are instead evidence about how a frozen adjudicator behaves: since the adjudicator neither receives nor remembers its record, a past ruling does not bind it. We also differ in the object of comparison: we compare separately developed machine adjudicators on identical disputes, control what an observer is shown about each, and measure forecast error separately from agreement with the source institution.

Evaluating LLM judges. LLM-as-a-judge evaluation asks whether a model's assessments of other models' outputs are reliable (Zheng et al., 2023). This literature has documented position bias and repetition instability (Shi et al., 2025), judicial bias and inconsistency (Hu et al., 2026), vulnerability to manipulation (Li et al., 2025), and difficulties in verifying rubrics for agentic outputs (Peng et al., 2026), and has developed procedures that decide when to trust a judgment and when to escalate (Jung et al., 2025; Badshah et al., 2026). Thakur et al. (2025) show that agreement summaries can conceal large differences between judges, so the insufficiency of agreement is already recognized. We add a different axis: not whether a judge's outputs are good or stable, but whether a party can anticipate them. Note that these properties are distinct: a judge can be consistently biased and easy to forecast, or stable across repeated calls and hard to forecast.

Predictive multiplicity. Our finding that adjudicators of similar accuracy decide over a third of contested disputes in opposite ways is an instance of predictive multiplicity: models with comparable performance can make conflicting individual predictions (Marx et al., 2020), and underspecified pipelines produce models that agree on held-out data but diverge in deployment (D'Amour et al., 2022). Our adjudicators are separately developed systems, not alternative solutions to one training problem, so we do not estimate a Rashomon set. Rather, we add an observer-dependent axis to multiplicity: disagreement measures whether two adjudicators differ, and legibility measures whether a party can tell how. At the institutional level, Kleinberg & Raghavan (2021) show that adopting a single, more accurate algorithm can still reduce welfare, which cautions against reading a favorable position on our map as a universal recommendation.

Simulatability and predicting models. Predicting a model's behavior, rather than solving its task, is an established evaluation paradigm (Doshi-Velez & Kim, 2017), used to test whether explanations help people anticipate models (Hase & Bansal, 2020; Hase et al., 2020). ALMANACS is the closest antecedent to our work: it predicts an LLM's answer probabilities on held-out, non-objective scenarios from earlier examples and explanations, and finds that explanations do not consistently help (Mills et al., 2025), much as rationales and rules do not help our observers. Poursabzi-Sangdeh et al. (2021) find that transparency can improve people's simulation of a model without improving their ability to correct it. Predictable AI argues that anticipating a system's behavior deserves attention alongside improving its performance (Zhou et al., 2025), and PredictaBoard evaluates assessors that predict whether a model will succeed on an instance (Pacchiardi et al., 2025). Other work predicts how an LLM judge will score a response (Ashley et al., 2026), and predicts unfamiliar AI agents from a few prior interactions, finding that target-adaptive learners outperform prompted LLMs (Shapira et al., 2026), in line with the agent-modeling literature (Albrecht & Stone, 2018; Rabinowitz et al., 2018). We note that we do not introduce probabilistic simulatability, but rather apply it to a setting where a party must commit to an adjudicator in advance, contrast it with agreement with an institution, connect forecast error to the margin that protects a party, and measure the effect of disclosing an adjudicator's rules and rulings.

Other meanings of legibility. The term has several established uses. In robotics, legible motion lets an observer infer an agent's goal (Dragan et al., 2013); in prover–verifier games, a legible solution can be checked by a weaker verifier (Kirchner et al., 2024); and recent work asks whether a reasoning trace shows which of its steps actually matter (Du et al., 2026). Adjudicative legibility, as we use it, denotes only whether an adjudicator's future rulings can be forecast. It does not recover hidden goals or certify that a justification is checkable or faithful.